Függőségi elemzésen alapuló magyar nyelvű keresőrendszer

Zsibrita János[1], Farkas Richárd[1], Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
{zsibrita,rfarkas,vinczev}@inf.u-szeged.hu

Kivonat A cikkben bemutatjuk webes keresőrendszerünket, mely függőségi elemzésen alapuló kereséseket tesz lehetővé magyar nyelvű szövegekben. A rendszer azokat a szókapcsolatokat adja vissza, ahol a keresett szó és annak bővítménye a keresőkifejezésben megadott nyelvtani viszonyban állnak egymással. Van mód a találatok szűkítésére is morfológiai, illetve szótőre vonatkozó jellemzőkre építve. A rendszer alapjait a magyarlanc morfológiai és szintaktikai elemző moduljai jelentik.
Kulcsszavak: keresés, szintaxis, morfológia, információkinyerés

1. Bevezetés

Az információkinyerés és -feldolgozás egyik fontos lépése a szövegek nyelvi előfeldolgozása, azaz a szövegek mondatokra, majd szavakra bontása, morfológiai elemzése és szófaji egyértelműsítése, illetve mély szintaktikai elemzése. A mély nyelvi jellegű adatok felhasználására épülő rendszerekkel általában pontosabb eredményeket kaphatunk, mint a felszíni jegyekre, puszta szóalakokra vagy szótövekre támaszkodó alkalmazások. Ezért a mély nyelvi elemzések kiaknázása igen hasznosnak bizonyulhat a nyelvtechnológiai alkalmazások terén, különösen a keresésen alapuló módszerek esetében.

Ebben a cikkben bemutatjuk függőségi elemzésre épülő keresőrendszerünket, melynek segítségével magyar nyelvű szövegekből nyerhetjük ki azokat a szókapcsolatokat, melyek a keresőkifejezésben meghatározott nyelvtani viszonyban állnak egymással. Tudomásunk szerint ez az első olyan, magyar nyelvű keresőrendszer, mely nagyméretű szöveges adatbázisokban képes szintaktikai alapú keresést végrehajtani, a szótövek kötelező meghatározása nélkül. A rendszerben lehetőség nyílik a találatok morfológiai alapon történő szűkítésére is. A továbbiakban részletesen bemutatjuk a keresőrendszert, majd példákkal illusztráljuk működését.

2. Kapcsolódó irodalom

A Nyelvtudományi Intézet gondozásában működő Nemzeti Korpuszportál[1] [1] felsorolja azokat a magyar nyelvű korpuszokat, amelyekhez rendelkezésre áll on-

[1] http://corpus.nytud.hu/nkp/

line kereső. Ezek közül most – mint általános szövegekben való keresőeszközöket – a Magyar Nemzeti Szövegtár [2] keresőjét és a Mazsola nevű eszközt [3] tekintjük át részletesebben.

A Magyar Nemzeti Szövegtár 1.0 változata [2] kb. 200 millió, 2.0 változata [4,5] közel egymilliárd szót tartalmaz. Mindkét változatában megtaláljuk minden egyes szó lemmáját és morfológiai elemzését, melyeket a keresésben is tudunk hasznosítani. A találatok konkordancia formában jelennek meg. Szókapcsolatokra is lehetséges keresni a keresett szó szűkebb környezetében előforduló más szavak vagy azok morfológiai jellemzőinek meghatározásával.

A Mazsola nevű eszközzel [3] a magyar igék bővítményszerkezetének feltérképezése válik lehetségessé. Szintén a Magyar Nemzeti Szövegtár [2] szövegeiben képes keresni, a szövegek morfológiai elemzését felhasználva. Segítségével megjeleníthető, hogy egy adott ige mellett milyen bővítmények jelenhetnek meg. Esetragok és névutók alapján, illetve a bővítmény szótöve alapján is lehetséges keresni az adatbázisban, majd a találatok gyakoriság szerint rangsorolva jelennek meg. A keresés egysége a mondat, azaz az igével egy mondatban szereplő bővítményeket ad vissza a kereső (melyek nem szükségszerűen szintaktikai vonzatai az adott igének).

A Mazsola mellett a jelenlegi munkában bemutatott keresőrendszerhez legközelebb a Szegedi Tudományegyetemen korábban kifejlesztett néprajzi kereső áll [6]. A kereső célja, hogy különféle néprajzi dokumentumokban egész mondatos kereséseket hajtson végre, azaz olyan dokumentumokat ad vissza, ahol a keresett ige és vonzatai a keresőkifejezésben megadott szintaktikai viszonyban állnak egymással. A háttéradatbázis magyar nyelvű hiedelmeket, táltosszövegeket, illetve meséket tartalmaz, összesen kb. 750 ezer szövegszóból áll. A keresőmondat függőségi elemzésére épülve a megtalált grammatikai relációknak és szótöveknek megfelelő illeszkedéseket keres a rendszer a szövegekben, morfológiai alapú, illetve szótövektől független keresésre azonban nem nyílik lehetőség.

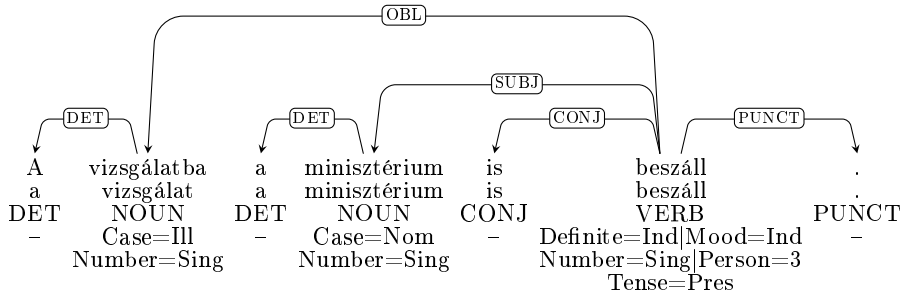
3. A keresőrendszer

Ebben a részben áttekintjük a keresőrendszer működési alapjait, illetve bemutatjuk röviden a mögöttes adatbázist.

3.1. A keresőrendszer működése

A magyar nyelv diskurzuskonfigurációs nyelv, azaz a bővítmények mondatbeli (szintaktikai) szerepére a szórend nincs hatással: a szintaktikailag összetartozó elemek nem feltétlenül szomszédosak, hanem előfordulhatnak egymástól távol is a mondaton belül. Ebből következően a magyar nyelvre nem optimálisak azok a keresési stratégiák, melyek pusztán a szövegkörnyezetet, azaz a keresett szó közvetlen környezetében található elemeket veszik figyelembe, szükség van szintaktikai információkra is.

A Javában implementált keresőrendszer az Elasticsearch[2] nyílt hozzáférésű eszközre épül. A keresések alapjául morfológiailag és szintaktikailag elemzett szövegek szolgálnak (lásd 3.2. rész). A morfológiai elemzés az univerzális morfológia magyarra adaptált elveinek [7] felel meg, míg a függőségi elemzésben a Szeged Dependencia Treebank [8] elveit követjük. Egy szövegszóhoz rendelkezésünkre áll annak lemmája, szófaja és részletes morfológiai elemzése, valamint a mondatbeli nyelvtani szerepe és az, hogy minek a bővítménye (azaz mi a szülő csomópontja a függőségi fában), lásd 1. ábra. A keresés során mindezen információt képesek vagyunk hasznosítani.



1. ábra: Morfológiai és szintaktikai annotáció.

A keresőrendszer alapvetően függőségi viszonyokra épül, azaz olyan szópárokat ad vissza találatként, amelyek között az adott szintaktikai reláció található (pl. alany-igei állítmány párok). A keresés során kötelező megadni a keresett függőségi viszonyt, továbbá lehetőség van a keresés szűkítésére más információk megadásával: mind a szülő, mind a gyermek csomópont esetében lehetséges azok lemmáját és/vagy szófaját, morfológiai jegyeit meghatározni.

Találatként olyan szópárokat kapunk vissza, amelyek között a megadott szintaktikai viszony szerepel, illetve minden további (szófajra, morfológiára, illetve lemmára vonatkozó) feltétel fennáll.

3.2. Az adatbázis

A keresőrendszer jelenleg két forrásból származó szövegekből képes visszaadni a meghatározott nyelvtani viszonyban álló szövegrészeket. Az egyik forrás a teljes Szeged Dependencia Treebank [8], melyben kézzel annotált morfológiai és szintaktikai elemzéseket találhatunk. Második forrásként az index.hu hírportálról töltöttünk le híreket (összesen 50,000 cikk) 2016 októberében és novemberében, majd azokat automatikusan elemeztük a magyarlanc 3.0 elemző lánc [9] segítségével. Az így kapott morfológiai és függőségi elemzések szolgálnak a keresések alapjául. Lehetőség van a találatok szűkítésére aszerint is, hogy melyik

[2] http://www.elastic.co/products/elasticsearch

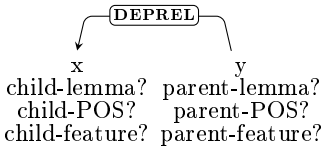
korpuszrészben szeretnénk keresni, így akár korpuszközi összehasonlításokat is végezhetünk.

A keresőrendszer mögött álló adatbázist folyamatosan bővítjük.

4. Keresési példák

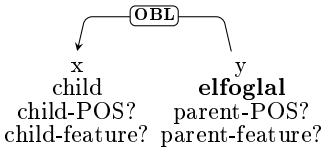
Az alábbiakban részletesebben is bemutatjuk, milyen keresési lehetőségeket biztosít a rendszer, továbbá arra is kitérünk, hogy a függőségi viszonyokon alapuló keresés milyen többletet jelent az egyszerű szó- vagy morfológiai alapú keresőkhöz képest.

A 2. ábrán láthatjuk a keresés sémáját. x és y jelöli a keresett szavakat (szóalakokat) – ezeket fogja visszaadni a keresőrendszer. Kötelező megadni a függőségi relációt (DEPREL), míg a kérdőjellel jelölt elemek opcionálisan megadhatók, akár egy, akár több elem is. Az alábbiakban ezekre mutatunk néhány példát.



2. ábra: A keresés sémája.

A szótövön alapuló kereséskor a keresett szó lemmáját kell megadnunk, illetve azt, hogy milyen szintaktikai viszonyt jelöl. Ha például arra vagyunk kíváncsiak, hogy milyen (nem alanyi, tárgyi és részeshatározói) vonzatai lehetnek az *elfoglal* igének, akkor a 3. ábrán látható keresést kell végrehajtanunk (vastaggal kiemelve a megadott elemeket):

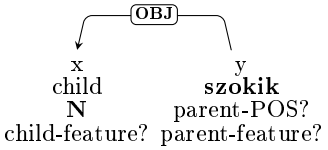


3. ábra: Keresés az *elfoglal* + OBL szerkezetre.

Lehetséges találatok például a következők:

pecsenyéjükkel/pecsenye@NOUN OBL elfoglalva/elfoglal@VERB
kihelyezésével/kihelyezés@NOUN OBL elfoglalva/elfoglal@VERB
Afganisztánban/Afganisztán@NOUN OBL elfoglalta/elfoglal@VERB

Egy másik példát nézve, a magyar *szokott* ige előfordulhat főigeként, illetve segédigeként is. Előbbi esetben jellemzően tárgyas igeként fordul elő, jelentése „hozzászokik", „megszok valamit", utóbbi esetben szokásos cselekvést jelöl, ilyenkor gyakran főnévi igenevet vonz, melynek szintén lehet tárgya. Ha arra vagyunk kíváncsiak, hogy mihez (milyen főnévhez) lehet hozzászokni, akkor az alábbi keresőkifejezést használhatjuk:



4. ábra: Keresés a *szokik* + OBJ szerkezetre.

A kapott találatok például a következők:

nagyapát/nagyapa@NOUN OBJ szoktam/szokik@VERB
világát/világ@NOUN OBJ szokták/szokik@VERB
vidéket/vidék@NOUN OBJ szokja/szokik@VERB

A fenti példák eredeti környezetéből kiderül, hogy valaminek/valakinek a megszokásáról esik szó éppen.

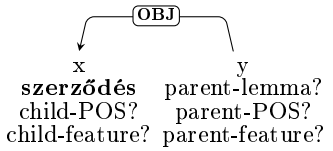
Ha nem áll rendelkezésre szintaktikai elemzés, akkor olyan keresőkifejezést írhatunk, ami a *szokik* ige közelében keres tárgyesetben álló főneveket. Ezzel téves találatokat is kaphatunk, például:

aranyórát/aranyóra@NOUN szokott/szokik@VERB

A fenti példát tartalmazó mondatból (*időnként el szokott lopni egy aranyórát*) jól látszik, hogy noha az *aranyórát* szó tényleg tárgyesetben áll, mégsem a *szokott* ige tárgya, hanem a *lopni* igéé, így nem állnak egymással közvetlen szintaktikai kapcsolatban.

A fentiek arra is rávilágítanak, hogy a szintaktikai információk felhasználásával pontosabb eredményeket kaphatunk, hiszen a fenti esetekben az igék és igenevek megfelelő vonzatait kapjuk meg, míg a pusztán szövegkörnyezetet (és morfológiát) felhasználó keresők a mondatban szereplő egyéb, nem az igenévhez tartozó vonzatokat is megjelenítenék.

A kereső arra is lehetőséget ad, hogy a vonzat oldaláról határozzuk meg a keresőkifejezést. Például ha arra vagyunk kíváncsiak, hogy a *szerződés* szó milyen szavaknak lehet a tárgya, akkor a következőképpen tehetjük meg:



5. ábra: Keresés a *szerződés* (OBJ) szerkezetre.

Lehetséges találatok a következők:

szerződést/szerződés@NOUN OBJ aláírták/aláír@VERB
Szerződést/szerződés@NOUN OBJ kötöttek/köt@VERB
szerződést/szerződés@NOUN OBJ írt/ír@VERB
szerződést/szerződés@NOUN OBJ bontott/bont@VERB
szerződést/szerződés@NOUN OBJ megtámadni/megtámad@VERB

A találatok gyakorisági mutatói arra is rámutatnak, hogy mik a magyar nyelvben gyakorta használatos szókapcsolatok, így akár a kollokációk vagy többszavas kifejezések megtalálásában és feltérképezésében is segítséget nyújthat a kereső.

5. Szintaktikai és szóalapú keresés

A magyar nyelvre eddig rendelkezésre álló keresők és lekérdezők többsége szóalapon működik, melyek a szóalakon kívül a szó lemmáját és morfológiai tulajdonságait képesek figyelembe venni a keresés során. Az általunk kifejlesztett rendszer több pontban is különbözik tőlük, melyeket az alábbiakban összegzünk:

- a keresés függőségi viszonyokon alapul, emellett a lemma és morfológiai információk is beépíthetők a keresőkifejezésbe,
- a bővítmény meghatározása is lehetséges, nem csak a szülő csomóponté,
- lexikális információ (a szóalak vagy szótő) meghatározása nélkül is tudunk keresni, csupán nyelvtani információk alapján,
- a keresés több találatot eredményez (nő a fedés), mivel a szintaktikai viszonyok figyelembevételével képes az egymástól távol eső, ám a lekérdezésnek megfelelő találatokat is visszaadni (pl. távoli függőségek),
- a keresés pontosabb találatokat eredményez (nő a pontosság), hiszen nem adja vissza azokat a szópárokat, amelyek lemmája és/vagy morfológiai elemzése megfelel a lekérdezésnek, ám egymással nem állnak szintaktikai kapcsolatban (pl. igenevek vonzatai).

6. Összegzés

A cikkben bemutattuk keresőrendszerünket, mely függőségi elemzésen alapuló kereséseket tesz lehetővé magyar nyelvű szövegekben. A rendszer azokat a

szókapcsolatokat adja vissza, ahol a keresett szó és annak bővítménye a keresőkifejezésben megadott nyelvtani viszonyban állnak egymással. A találatok pontosíthatók morfológiai, illetve szótőre vonatkozó jellemzők segítségével.

A szintaktikai alapokon nyugvó kereső felhasználhatósága többféle oldalról is jelentős. Egyrészt pontosabb találatokat ad, mint a pusztán szóalapon kereső rendszerek, így a magasabb rendű nyelvtechnológiai alkalmazások (pl. információkinyerés) is jobb eredményeket tudnak elérni. Másrészt korpusznyelvészeti és lexikológiai vizsgálatok során is lehetséges hasznosítani a keresőt. Továbbá akár a magyar nyelvtan, akár a magyar mint idegen nyelv oktatását is segítheti a rendszer.

A keresőrendszer mindenki által használható és szabadon elérhető a http://rgai.inf.u-szeged.hu/depsearch/ webcímen.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Hivatkozások

1. Sass, B.: Nyelvészeti szövegkeresők, Nemzeti Korpuszportál. Magyar Tudomány **7** (2016) 798–808
2. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria, European Language Resources Association (2002) 385–389
3. Sass, B.: The Verb Argument Browser. In Horák, A., Kopeček, I., Pala, K., Sojka, P., eds.: Proceedings of the 11th International Conference on Text, Speech and Dialogue, Berlin, Heidelberg, Springer Verlag (2008) 187–192
4. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
5. Oravecz, Cs., Sass, B., Váradi, T.: Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2015) 109–121
6. Zsibrita, J., Vincze, V.: Magyar nyelvű néprajzi keresőrendszer. In: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 361–367
7. Vincze, V., Simkó, K.I., Szántó, Zs., Farkas, R.: Universal Dependencies and Morphology for Hungarian – and on the Price of Universality (2017) Elfogadva az EACL 2017 konferenciára.
8. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771